人机信任校准的双途径：信任抑制与信任提升*

黄心语 李晔

（华中师范大学心理学院暨青少年网络心理与行为教育部重点实验室，武汉 430079）

摘 要 信任是人机成功合作的基础。但个体在人机交互中并不总是持有恰当的信任水平，也可能会出现信任偏差：过度信任和信任不足。信任偏差会妨碍人机合作，因此需要对信任进行校准。信任校准常常通过信任抑制与信任提升两条途径来实现。信任抑制聚焦于如何降低个体对机器人过高的信任水平，信任提升则侧重于如何提高个体对机器人较低的信任水平。未来研究可进一步优化校准效果评估的测量方法、揭示信任校准过程中以及信任校准后个体的认知变化机制、探索信任校准的边界条件以及个性化和精细化的信任校准策略，以期助推人机协作。

关键词 信任校准, 信任偏差, 信任抑制, 信任提升, 人机交互

1 前言

信任广泛存在于任何关系的建立与发展之中，如亲密关系(Rempel et al., 1985)、消费关系(Kwon et al., 2021)、组织关系(Meng & Berger, 2019)、医患关系(Petrocchi et al., 2019)等。它不仅是人际交往的重要因素，也是社会发展的润滑剂(乐国安, 韩振华, 2009)。随着机器人逐渐走进人们的生活，研究者们发现信任也存在于人机交互之中(Hoff & Bashir, 2015; Khavas, 2021)。本文结合前人的研究(高在峰 等, 2021; Lee & See, 2004; Mayer et al., 1995)，将人机信任定义为：个体在情境不确定或具有脆弱性时，对机器人能帮助己方实现目标或不会利用己方弱点时所持有的信心和心理预期。信任对于人机交互至关重要，它不仅是人类使用与接受算法的前提(Sanders et al., 2019)，也是人机合作的基础(Esterwood & Robert, 2021)。

本文主要关注人与智能机器人、算法、人工智能之间的交互，且以人与智能机器人之间的交互为主。以智能机器人为例，人机交互中，个体对智能机器人的信任水平可能过高，也可能过低，前者为过度信任(Over-trust)，后者为信任不足(Under-trust)。过度信任会导致人们对智能机器人不恰当的依赖和误用(Misuse)，信任不足则会导致弃用(Disuse)。信任不足和过度信任都会破坏人机交互系统的价值(Hancock et al., 2011)，因此个体需要在感知可靠性和实

收稿日期：2023-06-23

*国家自然科学基金面上项目(72371113; 71771102)资助。

通信作者：李晔，E-mail: liye@ccnu.edu.cn

际可靠性之间进行准确校准(Calibration)以保持恰当的信任水平(Madhavan & Wiegmann, 2007)。当个体拥有校准良好的信任时，他/她就知道何时应该信任机器人，何时不应该信任机器人(Ali et al., 2022)。人机信任往往通过两条途径进行校准：信任抑制(Trust dampening)与信任提升。信任抑制聚焦于如何降低个体不切实际的较高信任水平，信任提升侧重于如何提升个体过低的信任水平。需要注意的是，本文将提升个体过低信任水平的途径命名为“信任提升”，而非以往研究中经常使用的“信任修复”(Trust repair)。我们认为“信任修复”强调的是怎样改善个体在机器人出现信任违背(Trust violation)后的过低信任水平，它并没有把如何提升个体初始信任水平过低的情况包括在内。相比之下，“信任提升”能更好地囊括和反映该校准途径的内容。国外研究者们针对人机信任校准开展了大量研究(Alarcon et al., 2020; de Visser et al., 2020; Ososky et al., 2013)，考察了人机信任偏差的成因，并提出了相应的信任校准策略，但这些研究还比较分散，缺乏对该领域实证研究的系统整合与梳理；另外，目前针对人机信任校准策略的有效性尚存在争议，且以往研究大多只关注信任偏差的其中一方面(比如信任不足或过度信任)，忽略了从整体视角去整合信任偏差、信任校准有关研究的必要性与重要性。基于此，本文从人机交互中可能出现的信任偏差成因入手，梳理人机交互过程中机器人、个体本身、情境是怎样影响信任偏差的，以及如何通过信任提升与信任抑制两条途径校准人机信任、纠正信任偏差(见图 1)；本文也试图厘清人机信任校准策略的边界条件，并在此基础之上提出未来研究展望。

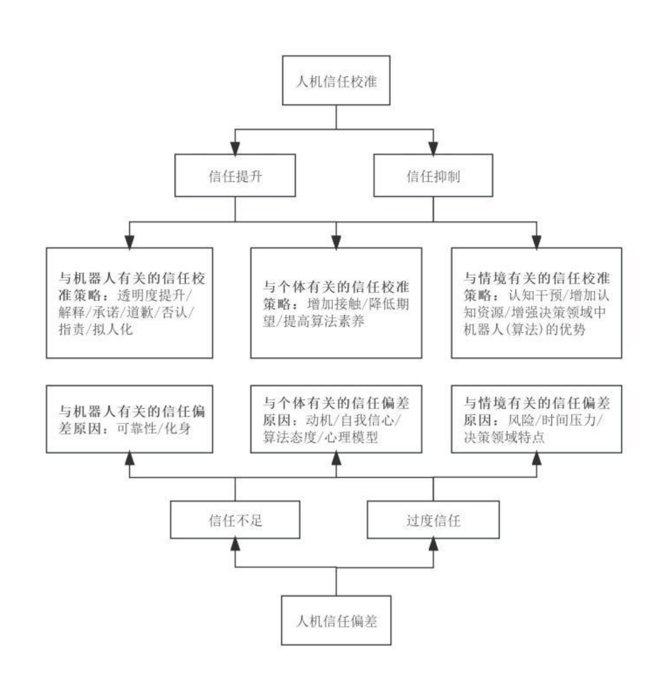


图 1 人机信任偏差之因以及校准策略

2 人机信任偏差

在本文中，我们将个体在人机交互中表现出来的过度信任和信任不足统称为人机信任偏差，即个体由于对机器人能力的错误估计导致信任偏离校准值。人机信任偏差会导致个体信任比人类更不可靠的算法，或不信任比人类更可靠的算法(Dzindolet et al., 2003)。

2.1 人机信任偏差的危害

过度信任往往出现在个体认为机器人具备人类没有的功能，或个体期望机器人能帮助他们降低风险的情境下(Borenstein et al., 2018; Parasuraman & Riley, 1997)。过度信任会直接导致个体高估机器人的能力，亦常常伴随有决策错误的风险，例如盲目地接受机器人智能体(Agent)提出的所有决策方案却不加考虑该决策是否合理(Khavas, 2021; Khavas et al., 2020)；过度信任有时甚至会对个体的生命造成威胁。Borenstein 等人(2018)发现尽管目前最先进的机器人外骨骼只能在低速慢走等有限条件下为运动残疾儿童提供一定的辅助功能，但是仍有很多运动残疾儿童的家长认为当他们的孩子进行某些风险运动(例如攀爬)时，外骨骼也可以保护他们的孩子不受伤害。这种过度信任机器(人)带来的危险同样出现在交通驾驶中：那些对于自动驾驶汽车高度信任的司机更容易在驾驶过程中出现打瞌睡的情况(Kundinger et al., 2019)，从而增加出现交通事故的可能性。

与过度信任相比，信任不足的危害较小，但是对于算法的信任不足往往会使个体倾向于低估算法能力(Parasuraman & Riley, 1997)，不能很好地利用算法，也无法享受使用算法所带来的好处(Ali et al., 2022)，最终在人机协作情境中恶化整体绩效，降低人机团队效率(Okamura & Yamada, 2020)。

2.2 人机信任偏差之因

2.2.1 与机器人有关的因素

可靠性。机器人本身与性能相关的因素，在人机交互之初对人机信任的影响很大(Robinette et al., 2017b)。以可靠性(Reliability)为例，它指代机器人性能具有前后一致性(Hancock et al., 2021)。一个可靠的机器人，应该具备可预测、性能稳定等特点。机器人的可靠性既可能诱发过度信任，也可能诱发信任不足(Shi et al., 2020)。具体来说，如果人们察觉到机器人的能力是可靠的、稳定不变的、可预测的，就极有可能放松对机器人的实时监测，表现出对机器人的过度信任；反之则容易出现信任不足。正如前文所述，机器人性能与可靠性密切相关，而人机交互之中机器人错误的出现则会诱发信任违背，导致信任方(个体)对受信方(机器人)的信任意向或信任信念降低(严瑜, 吴霞, 2016；Kim et al., 2009)。错误诱发信任违背的原因主要有二：一是错误会让人们怀疑算法的可靠性较低，进而造成信任水平下降(Alarcon et al., 2020; Correia et al., 2018; Lee & Moray, 1992)；二是人们往往对于算法错误这

类信息的敏感性较高，无法容忍算法出错，一旦算法出错就会直接弃用(Dietvorst et al., 2015)。另外，错误发生的频率、严重程度、数量也会影响信任水平的变化(Rossi et al., 2017)，错误发生的越频繁、越严重、数量越多，就会导致信任水平下降速度越快、幅度越大。除去机器人的明显错误以外，一些意料之外的、非预期行为同样也会导致信任违背。Lyons 等人(2023)发现当机器人行动路线偏离参与者原本设定的路线之后，参与者对机器人的信任感知和可信度感知均下降。

值得一提的是，针对错误是否会导致个体对机器人的信任下降，有研究者提出了相反的观点。例如 Sarkar 等人(2017)指出，错误不会影响参与者对机器人可信度的感知和后续的人机协作任务绩效，但他们同时也承认了在该实验中任务性质(任务困难且要求高)、犯错类型(仅涉及认知错误：给参与者错误指导，但不会妨碍其完成任务)等因素的影响。有趣的是，机器人的错误有时甚至被感知为可爱(Ragni et al., 2016)。当机器人犯错之后，人们会觉得它更具有人类相似性、更讨人喜欢(Mirnig et al., 2017; Salem et al., 2013)，一个完美的机器人反倒会看起来更不自然(Biswas & Murray, 2015)。这也印证了在人机互动中同样存在“出丑效应(Pratfall effect)”。在“剪刀石头布”游戏中，当机器人出现口头作弊(明明输了却声称自己赢了)或行为作弊(在看到对方出拳之后改变自己的原本答案)时，相比于无作弊行为的控制组机器人，人们在作弊条件下与机器人的社会互动明显增加，也更容易被机器人的作弊行为所逗乐，尽管他们主观上认为作弊是不公平的(Short et al., 2010)。

化身。人机交互中，化身(Embodiment)对信任也有一定影响，化身是指机器人的形态是实体或虚拟(van Maris et al., 2017)，主要划分为物理化身(Physical embodiment)和虚拟化身(Virtual embodiment)。物理化身指机器人在三维空间中具备一个有形的、物理的身体，能自由移动或操纵环境(Haring et al., 2021)；而虚拟化身(例如虚拟机器人)仅呈现在电子屏幕上，虽然拥有虚拟身体，但活动范围受限制。相较于虚拟呈现，人们更喜欢与物理呈现的机器人进行交互。物理化身会通过社会临场感(Social presence)影响信任，唤起个体对机器人的积极态度，将机器人视为社会行动者(Social actor)，进而对其作出社会化反应(Jung & Lee, 2004)。尤其当机器人的位置非常显眼，会无形增大个体依赖于它们的概率(Robinette et al., 2017a)。Bainbridge 等人(2011)发现，相对于远程呈现机器人，当机器人是以实体形态与参与者交互时，参与者对机器人不寻常命令的服从率会更高。在实体机器人形态条件下，尽管参与者都犹豫并感到困惑，但 22 名参与者中有 12 名还是听从机器人的指示将书本扔进了垃圾桶；相比之下，远程呈现机器人条件下仅有 2 或 3 名参与者服从了这个命令。需要说明的是，在该研究中服从机器人的命令被视为参与者信任机器人的表现。

2.2.2 与个体有关的因素

动机。过度信任可能出于人们社会惰性(Social loafing)的动机，即相较于自己单独工作，人机协作中个体付出的努力更少(Onnasch & Panayotidis, 2020; Parasuraman & Manzey, 2010)。当个体与机器人一起工作时，责任可能在个体和机器人之间分散，因此人机协作情境中更有可能出现“搭便车”效应(Dzindolet et al., 2002)。Cymek 等人(2023)的研究中发现，尽管单独工作组和与机器人协作组的参与者都自我报告在任务中投入了大量精力，但对比工作绩效发现单独工作组的参与者明显比机器人协作组的参与者任务绩效更高。Cymek 等人推测在实验阶段前四分之三的时间中，机器人协作组发现机器人的可靠性很高，因此在任务最后阶段放松警惕，未能及时察觉出机器人犯错。

自我信心。当个体的自我信心超过对自动化的信任时，个体更可能在人机交互中依靠自己；当个体对自身信心不足，则可能转向依赖于自动化(Lee & Moray, 1994)。在后一种情况下极易诱发过度信任，不仅仅因为人们认为算法较为权威，而人类的权力更小(Shank et al., 2021)、更弱势，也因为与人类决策相比，算法决策更为可靠、更准确(Mosier & Skitka, 1996)。Dijkstra(1999)的研究中，算法专家系统无论法律案件的具体情况如何总是认定罪犯有罪，参与者最后需要评估是否接受该系统的意见。研究结果发现，尽管参与者有更好的选择(例如听从人类律师辩护词的建议)，但他们最终更乐于听从算法专家系统的建议，即使建议是错误的。那些乐于听从算法专家系统的参与者对算法专家系统的评价更积极，权威依从得分更高。Xu 等人(2018)也发现相对于人类治疗师而言，人们更加信任机器人治疗师，并且伴随有过度信任的风险。

算法态度。算法态度是个体对算法的认知、情感和行为倾向的总和。算法欣赏(Algorithm appreciation)与算法厌恶(Algorithm aversion)就是两种典型的算法态度。算法欣赏会促使个体积极趋近算法决策，进而表现出对于算法的过度信任。Logg 等人(2019)发现，即使无法判断算法或人类决策的正确性，当参与者认为决策是来自算法而不是人类时，即使两者给出的决策内容实质是一样的，他们也更容易依赖算法，且这种算法欣赏效应具有跨主客观任务的一致性。个体对于机器人过高的信任也隐含了对机器人的性能期望(Lyons et al., 2020; Shin et al., 2020)。算法欣赏可能与高期望有关。期望越高，初始信任水平就会越高。一方面，高期望来自于对机器人外表的认知，例如机器人的拟人化(Anthropomorphism)会增强个体信任(van Pinxteren et al., 2019)；另一方面，高期望可能是缺失真实互动体验的结果。例如在一项研究中，当机器人在完成任务的同时报告“Q 值”(一串数字代码)，结果不管是否具备 AI 知识经验的参与者都认为 AI 更加可靠，认为越难以理解的 AI 越聪明(Ehsan et al., 2021)。

低信任水平也与算法厌恶息息相关。Chiarella 等人(2022)发现，两幅由同一位画家用不同色彩的颜料创作的绘画作品，仅仅通过操纵画作著者是人类还是 AI，就会影响人们对于画作的审美评分，具体表现为人们对“AI”作品的评分更低。算法厌恶可能是由于目前大众对机器人的实际接触较少，加之某些网络媒体对 AI 威胁的恐吓性报道和大肆宣扬 (例如 AI 会统治世界、未来将发生人机大战等) (Demir et al., 2019)，无形中加剧个体对机器人的负面态度，进而造成个体对机器人的消极印象。算法厌恶也可能是消极信任转移(Trust transfer)的后果(Okuoka et al., 2022)，如果个体之前对于计算机、手机等机械类产品有较糟糕的使用体验与经历，这种消极态度也会迁移到与算法有关的新兴产品上(Lee & Kolodge, 2020)。

心理模型。心理模型(Mental models)是经过组织的知识结构，亦是对工作环境的认知表征；人们使用心理模型来预测和解释他们周围世界的行为，并建构预期(杨正宇 等, 2003)。在人机交互的研究中，心理模型可以帮助个体更好地通过线索推断机器人的内在状态、并预测它的能力(Lee et al., 2005)。但是，由于心理模型是建立在个人经验的基础之上的，如果有新的经验发生，心理模型也会随之改变，因此个体之间的心理模型很可能各不相同(Müller et al., 2023)。人机信任校准的前提是人们能正确、全面、客观地看待机器人的优势与劣势，换句话说，个体需要具备恰当的心理模型以表征和理解机器人的能力。举例来讲，人机交互中，只有当机器人发出的信号被人类用户恰当解释时，人类才可以预测和理解机器人的行为(Breazeal, 2003)。因此，如果个体拥有对机器人恰当的心理模型，就能较好地校准信任，反之则会由于对机器人的能力估计错误而导致信任偏差(Ososky et al., 2013)。

2.2.3 与情境有关的因素

风险与时间压力。高风险条件或许会增大个体信任机器人的概率。Robinette 等人(2016)研究中，参与者在机器人的带领下前往会议室。机器人的带领路线有两种类型，一种是迂回的低效率路线，一种是不迂回的高效率路线。采用迂回路线带领参与者前往会议室的机器人被视为低能力的机器人。当参与者到达会议室后听到警报，需在一分钟之内立刻逃离这栋大楼。所有的参与者都跟随了机器人，甚至忽略了之前机器人的低能力。时间压力也会加剧过度信任，如果参与者感知到时间紧迫，更有可能向机器人寻求帮助，尽管之前已经观测到它出现过错误(Xu & Howard, 2018)。

决策领域特点。有研究者认为，与人类相比，人们对于算法是厌恶或欣赏的关键决定因素是该智能体背后的专业能力(Hou & Jung, 2021)。如果个体认为算法在某方面的专业能力不及人类，则有可能出现算法厌恶。譬如在医疗诊断中，人们在大多数情况下会认为人类决策优于算法决策，一方面是因为人类决策会让个体感到更有尊严，相比之下算法决策会给人

们带来非人化(Dehumanization)体验(Formosa et al., 2022)。而另一方面，当涉及到需要进行自我披露时，与机器人相比，人们也更愿意去信任人类，对人类的披露欲也更强(Barfield, 2021)。与此同时，决策领域的确定性程度也会影响人们是否使用算法。随着决策领域中不确定性的增加，人类和算法之间的绩效差异逐渐缩小，相对于“不完美”的人类犯错，人们更不能接受“完美”的算法犯错。因此，对错误敏感性降低的个体将倾向于依赖风险更高、误差更大的人类判断(Dietvorst & Bharti, 2020)。这种对于算法的偏见或是人们感知到相较于人类决策，算法决策更加不公平、不值得信赖，甚至算法犯错后还会引发更多的负面情绪(Lee, 2018)。

3 人机信任校准的途径

人机信任校准包括信任抑制与信任提升两条途径(见表 1)。信任抑制是指当机器人犯错却未被察觉或意外做出正确决策后，旨在降低个体不切实际的高信任水平的活动(de Visser et al., 2020)；信任提升定义为在初始交互时，亦或是信任违背后，旨在使信任方的较低信任信念和意愿更加积极的活动(Kim et al., 2004)。下面将分别从机器人、个体、环境三方面介绍对应的具体信任校准策略。

表 1 国外部分人机信任校准研究汇总

	信任校准途径	机器人类型	信任测量	研究的信任阶段	具体校准策略
Buçinca et al., 2021	信任抑制	虚拟人工智能	信任单维测量+行为测量	人工智能给出建议后	解释+认知强迫训练
Beller et al., 2013	信任抑制	自动化驾驶系统	信任单维测量+行为测量	每轮交互结束之后	呈现不确定性信息
Wang et al., 2018	信任抑制	机械机器人/动物机器人	人际信任量表改编	每轮交互结束之后	解释+化身
Lyons et al., 2023	信任提升	机械机器人	人际信任量表改编	机器人出现非预期行为前后	承认自己的过失+解释非预期行为出现的原因
Kim & Song, 2021	信任提升	虚拟智能代理	人-机器信任问卷改编+行为信任测量(依从性)	每轮交互结束之后	拟人化+道歉
Sebo et al., 2019	信任提升	类人机器人 NAO	人际信任问卷改编+行为测量	人机交互之后	否认+道歉

3.1 与机器人有关的信任校准策略

透明度提升。提升机器人透明度既可以用于降低过度信任(de Visser et al., 2020)，同时也可以用于提升信任不足(Lyons et al., 2017)。但总体而言，透明度常常用于纠正过度信任。透

明度包括向用户提供关于模型如何工作的相关信息(Bhatt et al., 2020)，以帮助他们理解该系统(Seong & Bisantz, 2008)。算法需要具备可理解性(Understandability)，让用户了解算法内部的底层运行机制，信任并正确地使用算法系统。透明度还包括公开机器人的内在言语(Inner speech)，把机器人做出决策的推理过程、动机过程、目标和行动计划展现给用户(Chen et al., 2018; Geraci et al., 2021)，从而抑制信任。机器人也可以及时向用户提供有关性能的反馈，例如通过语音的方式传达它对于决策正确与否的不确定性(Okamura & Yamada, 2020)。正如前文所述，可预测性是机器人可靠性的重要组成部分之一。当机器人性能不稳定、不可预测时，个体就无法对机器人的可靠性进行准确评估。因此，通过向用户传达(性能)不确定性，暗示机器人可能在后续的交互过程中出现性能下降的情况，有利于抑制过高信任。Beller 等人(2013)通过驾驶模拟任务检验不确定性(即自动驾驶汽车在性能不确定的情况下出现一个带有迟疑表情的 emoji)在信任校准中的作用。研究显示，与控制组相比，不确定组能降低用户对于自动驾驶汽车的依赖，提醒用户为自动化故障做好准备，并能促使用户更主动、更快地处理自动化驾驶汽车的故障。不确定组的参与者也更能在驾驶任务中集中注意力，更不容易被其他无关刺激所干扰。该研究结果与 Kunze 等人(2019)基本一致，不确定性反馈有助于用户调整注意力的分配，进而校准信任。但不确定性信息该如何呈现则需要设计师仔细斟酌，因不确定性呈现虽会有利于信任校准，但可能诱发更高的工作负载，从而降低任务绩效(Kunze et al., 2019)。

显示信心指数亦是提升透明度的策略之一(de Visser et al., 2020)。信心指数即 AI 作出正确决策的概率，理论上人们应该在 AI 报告信心指数高的情况下依赖 AI，在信心指数低的情况下依赖自己的判断。McGuirl 和 Sarter(2006)发现，如果自动化能提供动态更新的信心指数将有助于飞行员在任务分配和是否遵守自动化系统的建议等方面做出更好的决策，对系统准确性的估计也更加精准。类似的，在人机交互中，如果机器人能经常向用户提供它对完成某件任务的信心指数，个体或许就可以根据该指数合理分配任务给机器人。

解释。可解释的 AI(Explainable AI , XAI)是人们正确校准信任的必要组成部分(Adadi & Berrada, 2018)，亦是信任抑制的主要策略之一(Buçinca et al., 2021)。XAI 需要为用户提供有意义解释，同时也可向用户索要解释(de Visser et al., 2020)，它通过让用户了解 AI 的决策过程，以期 AI 给出错误决策时能被用户准确识别并拒绝。人们之所以对机器人的期望过高，机器人的“黑匣子”(Black box)属性有可能是其中的重要因素。如果在人机交互之前能及时打开“黑匣子”，或许能降低个体过高的信任水平，并帮助个体建立起对于机器人的良好心理模型。Wang(2018)等人发现，解释有利于校准参与者的信任，从而帮助参与者更好地做出

决策。当机器人没有给出任何解释时，参与者会过度信任机器人进而导致决策失误。相比之下，当机器人提供解释时，参与者对机器人的依从率就降低了。除此之外，适当地传达机器人的局限也可以进一步纠正个体对于机器人高可靠性的期待，例如机器人明确地告知用户自己能够执行的任务和功能范围，从而避免个体出现滥用机器人的情况(de Visser et al., 2020)。

与此同时，人机交互中当机器人出现错误时，适当的解释有利于个体更好地了解错误发生的机制，并通过给出相关证据来加强解释的说服力，从而提升信任。解释包括说明错误发生的原因(Correia et al., 2018)，承认错误事件的发生并给出一个能够推断出因果关系的理由(Bhatt et al., 2020; Lyons et al., 2023)、提出解决问题的办法(Hald et al ., 2021)。解释需要与用户的知识背景相匹配(Adadi & Berrada, 2018; Kim & Hinds, 2006)，如果提供一个过于专业化的术语解释，反而会让用户一头雾水，并降低机器人的透明度。然而，解释有时也会弄巧成拙(Papenmeier et al., 2019)，它是否有效可能跟该错误导致的后果是否严重有关。在 Correia 等人(2018)的研究中，机器人与参与者合作完成拼凑七巧板的任务。当机器人突然出现语音故障导致游戏暂停时，只有当参与者和机器人可以继续完成剩下的任务时，机器人的解释才有效，当需要重启游戏时，解释就无用了，甚至在这种情形下的解释还会导致参与者的信任水平下降。

承诺。承诺适用于正直型违背或能力型违背，前者侧重于被信任方(Trustee)因为他/她的诚实品质问题而造成信任方(Trustor)的信任下降，后者则强调被信任方因能力不足而没有达到信任方的期待造成信任下降(严瑜, 吴霞, 2016)。承诺不仅包括了信任违背之后机器人给予人类的承诺，还包括了人类事先给予机器人的承诺。针对前一类，Esterwood 和 Robert(2022)发现，当个体之前对机器人的积极态度高时，承诺对修复信任最有效。承诺通过保证个体对机器人所持的态度是正确的来强化个体的积极态度，进而减少认知失调，更有利于信任修复。针对后一类，Sebo 等人(2019)发现，如果人类与机器人在交互前事先进行了互惠性承诺，也就是以不伤害对方的前提下公平竞争，即使在任务中机器人耍赖并欺骗了参与者，事后相较于没有做出事前互惠承诺的个体，参与者仍报告了对机器人较高的信任。

道歉。道歉是人机交互中修复信任最常见的方法，它适用于能力型信任违背(Quinn, 2018)。道歉被定义为承认自己因信任违背行为所带来的责任，并表达遗憾(Kim et al., 2004)。道歉常常会跟归因相联系。例如 Kim 和 Song(2021)发现，当发生信任违背后，相比于使用外部归因道歉策略，使用内部归因道歉策略的类人化虚拟智能体更能修复信任；该结果恰恰与类机器虚拟智能体相反。当机器人表达出类似于人类的情绪，例如遗憾时，相较于没有表现出遗憾的机器人，参与者对其信任度急剧增加；当道歉既包括遗憾的语言表达，又包括解

释时，信任水平的增长尤为明显(Kox et al., 2021)。同时，道歉时机也很重要。Robinette 等人(2015)通过模拟火灾危机情境发现，在信任违背之后机器人的立刻道歉和解释不能有效地修复信任，而在危机时刻机器人同样使用道歉和承诺时，大部分的参与者会选择重新跟随机器人前往紧急出口。但 Quinn(2018)也质疑道歉可能会因为机器人反复表达内疚和感知的低真诚而降低信任修复的有效性。

否认。否认对于修复正直型违背十分有效(Sebo et al., 2019)。否认常包括否认外部因果关系，既不承认任何责任，也不表示遗憾(Kim et al., 2004)。否认给予了信任违背者一个机会去反驳与质疑，而不是单纯地承认错误。同时，它也表明一种没有必要去改正行为的意向，这可能会导致个体对违背者未来信任行为的担忧(Kim et al., 2004)。但相对于道歉策略直接将机器人的失败暴露于人类面前，当个体处于在高工作负载条件，且无法验证机器人的正直性或无法厘清故障原因时，否认可能是一种更安全的修复策略(Quinn, 2018)。有趣的是，当机器人出现正直型违背后给予否认，虽然后续参与者报告的信任水平与其他条件无差异，但是有 60%的参与者会选择在实验中对机器人进行报复(Sebo et al., 2019)。

指责。指责是信任修复中风险较大的一种策略。与道歉相类似，指责通常也会涉及到归因问题，且最好让机器人在引入指责归因时，将任务失败归结于机器人自己内部的原因(Groom et al., 2010)，而不是外部原因(算法设计师、第三方算法、人类同伴)。信任违背之后，相比于外部指责，通过内部指责进行解释虽然在行为信任上不存在显著的差异，但是却能让参与者感知到更强的正直性和仁爱性(Jensen et al., 2019)。同样，也并不是所有的指责都有效，如果这类指责仅强调是自己的错误而不去指出发生错误的原因，那么这个时候机器人的指责归因的引入亦会导致参与者的信任感知下降(Kaniarasu & Steinfeld, 2014)，因为一个指责他人(尤其是参与者本身)的机器人会让参与者感到愤怒，但一个自怨自艾的机器人同样也让人觉得不被信任，尽管它们很诚实地指出了自己的错误。

拟人化。拟人化是将人类特征、动机、意向或心理状态赋予非人对象的心理过程或者个体差异(许丽颖 等, 2017; Epley et al., 2007)，它可以用于提升人机信任。因算法常被知觉为冷冰冰、缺乏温暖和体验性，如果能在算法中添加一些与人类相似的高情感特征，比如使用女性机器人(高温暖与体验性的代表)，或许会缓解人们对机器人去人性化的感知(Borau et al., 2021)。Toader 等人(2019)证实，与男性聊天机器人相比，跟女性聊天机器人互动的参与者对于个人信息的披露意愿更强，社交感知和服务满意度也明显更高。由于拟人化需要将人类特征投射到机器人上，因此，拟人化可能会让参与者产生“机器人就像人类一样容易出错”的认知(Aroyo et al., 2021)，进而增加信任弹性(Trust resilience)，帮助个体形成有关于机器人的

心理模型(Ososky et al., 2013)，减缓错误后参与者的信任下降速度(de Visser et al., 2016)。在一项研究中，机器人的主要任务是给参与者递四个鸡蛋让参与者能顺利制作鸡蛋卷。相比于另外两个不能交流的机器人，一个能交流、能表达自己情绪(例如掉落鸡蛋之后做出委屈的表情)、但偶尔犯错(运输鸡蛋过程不慎掉落一颗鸡蛋)的机器人更受参与者喜爱，甚至当它犯错之后，参与者对其的信任程度仍不亚于一个效率高(不掉落鸡蛋)但沉默的机器人(Hamacher et al., 2016)。

然而，拟人化也可能诱发过度信任。用户可能会过度信任拟人化程度高的机器人，因为高拟人化机器人往往会被知觉为更可靠、更仁爱、更诚实，导致用户对机器人产生一种错误的熟悉感，从而诱发对其类人的预期(Wagner et al., 2018)。因此，对于初始信任较高的个体，降低机器人的拟人化特征也是抑制信任的方法之一。

3.2 与个体有关的信任校准策略

增加接触。增加接触在一定程度上能改变个体对机器人的态度。在信任提升方面，研究证实曝光效应(Exposure effect)同样存在与人机交互之中(Jessup et al., 2020; Wullenkord et al., 2016)。与机器人的面对面互动会减弱人们对机器人的警惕(Haring et al., 2013)，减少不确定性和风险感知(Kraus, Scholz, Messner, et al., 2020)，增进个体对于机器人的好感，提升人机初始信任。有趣的是，仅仅通过帮助机器人按下一个按钮，相较于没有按钮的参与者来说，参与者对于机器人的信任水平也会更高(Ullman & Malle, 2017)。总而言之，与机器人的实际接触可能会降低个体对机器人的负面偏见和焦虑，纠正个体过去对于机器人可能构成威胁的不恰当认知，最终提高个体未来与机器人继续接触的意图(Wullenkord et al., 2016)。

接触也可以在一定程度上改变个体对于算法不恰当的欣赏，从而降低过高信任。人机交互经验已被证明与自动化依赖息息相关(Goddard et al., 2012)。Haring 等人(2013)的研究中，与机器人交互前个体可能认为机器人较为聪明，但是当真正与机器人交互之后，参与者对机器人的拟人化感知、智力感知均有降低。该研究的结果在 Wullenkord 等人(2016)的研究中得到了重复：参与者在人机交互前认为机器人的情绪体验性较强，但是真正与机器人互动之后，他们便会逐步意识到机器人不是很先进，可以体验到的情绪也比他们想象的要少；相比之下，控制组(即没有与机器人交互过)的参与者仍然秉持着机器人能力较强的观点。与机器人接触过的实际经验使个体对机器人能力认知开始趋于正常化(Sanders et al., 2017)，从而达到校准信任的效果。

降低期望。降低期望是人机信任抑制的方法之一。信任的动态变化特点促使个体在交互过程中根据新信息不断校准对机器人性能的期望，因此个体对于机器人的认知也会随着与机

器人交互的深入而不断更新 (Kraus, Scholz, Stiegemeier, et al., 2020)。Pop 等人(2015)发现，对自动化具有高期望的用户虽然对于自动化可靠性的变化更为敏感，但却不一定会具有较好的信任校准能力。当自动化能力提高时，用户的信任校准较好，而当自动化能力降低时，信任校准较差。因此，如果个体对机器人的期望较高，事先预警(Forewarning)是比较有效的方法。通过预先警告该任务难度、提醒用户自己可能在该任务中表现不佳(de Visser et al., 2020)，进而帮助个体重新设定他们的期望(Lee et al., 2010)。

提高算法素养。算法素养(Algorithm literacy)主要包括四个方面：(1)用户了解 App 以及平台算法是如何被使用的(2)用户知道算法如何运行(3)用户能够批判性地评估算法做出的决策(4)用户能有效处理算法运行过程中出现的问题(Dogruel et al., 2022)。如果个体具备良好的算法素养就可以与机器人顺利交互，并从机器人的解释中提取新的知识，进而改善心理模型(Naiseh et al., 2021)。算法素养可以通过学习提升，例如在机器人用户手册中强调过度信任的风险，并列举过度信任机器人的优劣；机器人运行商可以开发机器人培训相关的课程，帮助人们正确地了解机器人(Aroyo et al., 2021)，从而提升个体的自我信心，降低对于机器人的高依从性。其次，用户同样可以通过自我学习可以提高学习能力，并更新他们对 AI 的知识，以最有效的方式来校准信任。

3.3 与情境有关的信任校准策略

认知干预、增加认知资源。情境的特点会影响认知资源负荷。首先，在高风险与时间压力下的个体往往会出现认知资源过载的情况。根据认知负荷理论(Cognitive load theory)，人的工作记忆能力是有限的，而认知负荷又分为内在认知负荷和外在认知负荷。内在认知负荷主要是由学习任务产生的，而外部认知负荷则来自于与学习任务无关的其他来源，例如环境(Sweller, 2011)。从该理论出发，人机交互过程中个体的认知负荷较高容易导致个体无法准确识别自动化错误；而不断监督自动化运行情况可能会让个体产生与任务无关的内在认知负荷(Lyell & Coiera, 2017)，认知资源越少，个体就越容易出现过度信任算法的倾向(Chien et al., 2016)。因此优化人机交互环境或许有利于提高认知资源利用率并抑制信任，例如简化个体的用户界面(Naiseh et al., 2023)，以清晰可理解的方式提供指示(Wickens, 1995)。

其次，快速决策情境中也容易受到认知启发式的影响。在此基础之上，Buçinca 等人(2021)提出了相应的认知干预策略。他们以认知的双重加工理论为切入点，指出人们的认知过程包括双重系统，第一重系统是启发性思维阶段(包括启发式和心理捷径)，即人们为减少认知资源损耗而通过单一线索作出判断与决策(Tam & Ho, 2005)；第二重系统是分析性思维阶段。总体来说，人们的大多数日常决策都是通过启发性思维完成的，分析性思维因其触发缓慢、

需要较多认知资源而很少被激活。他们通过训练参与者进行认知强迫(Cognitive forcing)，要求参与者(1)先于 AI 做出决定(2)通过增加 AI 给出建议的时间去放缓决策过程(3)让参与者选择是否以及何时查看 AI 建议。研究结果表明，认知干预增加了参与者进行分析性思维的认知动机，进一步减少了参与者对 AI 的过度依赖。

增强决策领域中机器人(算法)的优势。Hou 和 Jung(2021)认为，人类并不是一味地偏向于算法或人类决策，个体实质偏向的是算法或人类背后的专业能力。适度在算法决策背后注入专家力量有利于改善个体原先的消极信任态度。除此之外，在不同的任务领域匹配不同外表的机器人可以帮助个体更好地接纳机器人，比如享乐主导的服务环境中，个体对像儿童的高热情服务机器人表现出更高的偏好，而在功利主导的服务环境中，他们对像成人的高能力服务机器人表现出更高的偏好(Liu et al., 2022)。人们通常厌恶在一些较为主观的任务中使用算法，但如果能在这些看似主观的任务中强调一些可以用客观事实解释的部分，也可以降低主观任务的“主观性”，使参与者可以更好地接受算法决策。比如告知参与者歌曲的推荐(主观任务)一定程度上也可以由个体的人格特质(客观性)所决定，这时个体对于算法决策的信任会增强(Castelo et al., 2019)。

4 未来研究展望

人机交互已经逐渐渗透到日常生活的方方面面，而信任对于保持团队凝聚力至关重要(Perkins et al., 2021)。然而，信任只有维持在恰当的水平才能促进有效的人机合作。一旦信任出现过高或过低的情况，就会对人机合作造成一定的威胁，因此需要对信任进行校准。目前针对人机信任校准虽然已经有较多研究成果的积累，但是仍存在不足和值得改进的地方。

4.1 优化校准效果评估的测量方法

第一，从测量方法上看，在人机信任领域已经有研究者开始采用内隐方式考察个体对自动化的信任(Merritt et al., 2013)，但仍限于个体信任校准之前对自动化/机器人的信任测量，未涉及到个体在人机交互中出现了信任偏差并进行信任校准之后的后续内隐信任测量。第二，校准信任之后研究者们往往以信任量表得分或信任行为等外显信任指标作为衡量信任修复/抑制效果的主要方法。我们认为，校准信任之后不仅要关注个体的外显信任态度，同时也要关注个体的内隐信任态度以更好地检验校准策略的有效性与实用性。以信任抑制为例，未来不仅可以通过信任量表检验抑制策略是否有效，亦可通过内隐联想测验探讨信任抑制后个体的内隐信任水平是否也有降低，对比信任抑制策略在降低外显信任和内隐信任之间的差异。

4.2 揭示信任校准的认知神经过程

以往人机信任的有关研究大多停留在行为实验上，目前已经有部分研究者开始从认知神经的视角去关注人机信任(Eloy et al., 2022; Oh et al., 2020; Walker et al., 2019; Yen & Chiang, 2021)，例如机器人出错后，个体的内侧和右侧背外侧前额叶皮层内观察到神经激活增加，大脑的功能连接强度降低(Hopko & Mehta, 2022)，前扣带皮层的负波增加(de Visser et al., 2018)。信任是一个连续的过程，信任的建立、增长、受损和消失会对信任关系中每个成员的目前及未来的行为方式产生强大而持久的影响(Hancock et al., 2011)。同样的，完整的信任校准周期往往会经历信任建立-信任增长/受损-信任校准三个阶段。以往人机信任的认知神经研究比较关注前两个阶段，然而在人机信任校准过程中，尤其是对个体的偏差信任水平进行信任提升或信任抑制之后，涉及到哪些认知神经过程还鲜有研究，而这一部分恰恰对于人机信任校准极为重要，不仅可以揭示人机信任校准的生理机制，也可为后续的信任校准策略优化提供思路与借鉴。未来研究者可利用生理指标实时、全程监测个体从信任建立之初到信任校准后的认知神经活动变化过程，进一步从生理层面揭示个体的信任动态化发展。

4.3 结合信任发展阶段对信任校准策略进行精细化研究

信任本身呈现动态发展变化的特点，以往对于信任校准的研究大多是基于静态的横断面研究，仅考察了在当前阶段中如何提升或抑制个体的信任水平，并未以动态的视角考察信任水平发展变化的影响因素。以信任过高为例，人机交互之前个体可能先入为主地持有对于算法的消极态度，并认为算法的能力较低。如果在后续交互过程中个体感知到了算法的可靠性，这种原先对于算法低能力的预期就会被打破，期望落差就有可能更会促使人们趋近算法，并认为算法更加值得信任(Washburn et al., 2020)。Filiz 等人(2021)发现，在 40 轮的股价预测实验中，参与者可以选择相信自己或相信算法，但最后的报酬会与预测正确率挂钩。虽然有的参与者刚开始选择相信自己，但随着他们发现自己的预测准确率低于算法时，他们也会逐渐地选择相信算法。该结果在另一项研究中得到了重复：当机器人记者所撰写的新闻质量超过了人们的预先预期时，这种对于机器人记者期望的积极失验(Positive disconfirmation)会让人们更愿意接受机器人记者，并且也会更加满意(Kim & Kim, 2021)。在这种情况下人们对机器人持有的高信任水平，与个体在人机交互过程中的逐渐积累起来的高信任水平可能有所差异，因此应该采用不同的信任抑制策略，但以往研究并未加以区分。未来可进一步针对人机信任偏差产生的不同原因，分别比较不同的校准策略的有效性，探索最适宜某种信任偏差的校准策略。

4.4 探讨信任校准的边界条件

首先，目前人机信任的研究几乎都在考察人对类人机器人、机械化机器人信任水平的发

展变化，而较少关注动物型机器人在信任校准中的作用，尤其是“萌萌的”动物机器人，可能会诱发个体对其天真、善良等特质的自动推断，唤起人们的积极情绪(许丽颖 等, 2019)。一个有着圆圆的、大大的眼睛的娃娃脸机器人也可能被认为更可信(Song & Luximon, 2020)。动物型机器人比机械化机器人更受人喜爱(Li et al., 2010)。人类独特性或是人们对机器人的初始信任水平较低的原因之一，既然娃娃脸能够降低人们对其威胁性的评判(许丽颖 等, 2019)，那么萌萌的机器人也许可以改变个体的偏见从而提高初始信任水平；同样发生信任违背之后，萌萌的动物型机器人可能也会使得信任水平下降的更慢，更容易被修复。对于信任抑制来讲，与类人机器人相比，动物型机器人能较好地通过熟悉度降低个体期望，同时避免了类人机器人设计中可能存在的种族偏见等问题(Löffler et al., 2020)，进而降低个体过高的信任；人们也可能会因为类人机器人的外表对其产生不恰当的认知，反观动物机器人或许会降低个体对其心理模型的推论，从而抑制信任。未来可进一步对比类人机器人与动物型机器人在信任校准方面的作用，但需要注意的是，动物型机器人最好具备较高或较低的动物相似度，否则容易出现恐怖谷效应(Löffler et al., 2020)。

其次，目前人机信任领域已经有研究者开始关注个体在群体之中、而不是单独与机器人交互时信任水平的变化发展(Montague & Xu, 2012; Montague et al., 2014; Xu & Montague, 2013)。例如 Martinez 等人(2023)考察了个体单独和作为群体成员(2~3 人)在点餐前、点餐后，从机器人那里取到外卖三个阶段中对于送餐机器人的信任与接受度，结果发现随着与机器人接触增多，个体的参与者对于机器人的信任水平也在逐渐增长，但是群体中参与者的信任却并未随着接触的增加而增长，反而有更多的变异。也有研究者探索了从众在人机信任中的影响，发现相较于与直接和机器人沟通从而建立起对于机器人的信任，人们更喜欢听从其他人对于机器人的评价并在此基础之上作出信任判断(Volante et al., 2019)。这两项研究初步探索了个体在群体中可能会出现的信任水平变化，但未涉及如何校准群体中个体的信任水平。未来可通过跨文化的方式比较中西方参与者在群体之间的人机信任水平差异并进一步考察如何在群体内进行信任校准，亦可比较个体信任偏差与群体信任偏差的差异与共性，探索适宜群体信任偏差校准的策略。

最后，信任校准能否成功其实也取决于个体的因素，校准策略的有效性可能存在个体差异。例如 Lee 等人(2010)发现，对于不同服务导向的个体可以采用不同的信任修复策略，持有关系导向的参与者更喜欢道歉这类信任修复策略，而持有功利导向的参与者更喜欢赔偿这类实质行为上的信任修复策略。未来可进一步根据用户自身的特点，在人机交互过程中捕捉不同个体与信任相关的行为进行建模(Pynadath et al., 2019)，以个性化的方式校准信任。

致谢：感谢西南科技大学赵文老师为英文摘要润色，感谢两位外审专家和编委为本文修改提出了宝贵的修改意见和建议。

参考文献

高在峰，李文敏，梁佳文，潘晗希，许为，沈模卫. (2021). 自动驾驶车中的人机信任. *心理科学进展, 29*(12), 2172–2183.

许丽颖，喻丰，邬家骅，韩婷婷，赵靓. (2017). 拟人化：从“它”到“他”. *心理科学进展, 25*(11),1942–1954.

许丽颖，喻丰，周爱钦，杨沈龙，丁晓军.(2019).萌:感知与后效. *心理科学进展, 27*(4), 689–699.

严瑜，吴霞.(2016).从信任违背到信任修复:道德情绪的作用机制. *心理科学进展, 24*(4), 633–642.

杨正宇，王重鸣，谢小云. (2003). 团队共享心理模型研究新进展. *人类工效学, 9*(3), 34–37.

乐国安，韩振华. (2009). 信任的心理学研究与展望. *西南大学学报(社会科学版), 35*(2), 1–5.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020, September). Trust repair in performance, process, and purpose factors of human-robot trust. *In 2020 IEEE International Conference on Human-Machine Systems (ICHMS)* (pp. 1–6). Rome, Italy.

Ali, A., Tilbury, D. M., & Jr, L. R. (2022). Considerations for task allocation in human-robot teams. *arXiv preprint arXiv:2210.03259.*

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., ... & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics, 12*(1), 423–436.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics, 3*, 41−52.

Barfield, J. K. (2021, August). Self-disclosure of personal information, robot appearance, and robot trustworthiness. *In 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 67–72).Vancouver, BC, Canada.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction: An approach using automation uncertainty. *Human Factors, 55*(6), 1130–1141.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020, January). Explainable machine learning in deployment. *In Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 648–657). https://doi.org/10.1145/3351095.3375624.

Biswas, M., & Murray, J. C. (2015, September). Towards an imperfect robot for long-term companionship: Case studies using cognitive biases. *In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5978–5983). Hamburg, Germany.

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing, 38*(7), 1052–1068.

Borenstein, J., Wagner, A. R., & Howard, A. (2018). Overtrust of pediatric health-care robots: A preliminary survey of parent perspectives. *IEEE Robotics & Automation Magazine, 25*(1), 46–54.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems, 42*, 167–175.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*, CSCW1, 1–21.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825.

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in*

Ergonomics Science, 19(3), 259–282.

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards Artificial Intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior, 137*, 107406.

Chien, S. Y., Lewis, M., Sycara, K., Liu, J. S., & Kumru, A. (2016, October). Influence of cultural factors in dynamic trust in automation. *In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2884–2889). Budapest, Hungary.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018, July). Exploring the impact of fault justification in human-robot trust. *In Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 507–513). Stockholm, Sweden.

Cymek, D. H., Truckenbrodt, A., & Onnasch, L. (2023). Lean back or lean in? Exploring social loafing in human-robot teams. *Frontiers in Robotics and AI, 10*, 1249252. doi: 10.3389/frobt.2023.1249252.

Demir, K. A., Döven, G., & Sezen, B. (2019). Industry 5.0 and human-robot co-working. *Procedia Computer Science, 158*, 688–695.

de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in Human Neuroscience, 12*, 309.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics, 12*(2), 459−478.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science, 31*(10), 1302−1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114−126.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology, 18*(6), 399−411.

Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures, 16*(2), 115–133.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697−718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79−94.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., & Riedl, M. O. (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509.*

Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., & Hirshfield, L. (2022). Using fNIRS to identify transparency-and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads. *Frontiers in Neuroergonomics, 3*, 838625.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114, 864–886.

Esterwood, C., & Robert, L. P. (2021, August). Do you still trust me? Human-robot trust repair strategies. *Proceedings of 30th IEEE International Conference on Robot and Human Interactive*

Communication.Vancouver, BC, Canada.

Esterwood, C., & Robert, L. P. (2022, March). Having the right attitude: How attitude impacts trust repair in human-robot interaction. *In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 332–341). Sapporo, Japan.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance, 31*, 100524.

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. *Computers in Human Behavior, 133*, 107296.

Geraci, A., D'Amico, A., Pipitone, A., Seidita, V., & Chella, A. (2021). Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions. *Frontiers in Robotics and AI, 8*, 620026.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.

Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010, March). Critic, compatriot, or chump? Responses to robot blame attribution. I*n 2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 211–217). IEEE.

Hald, K., Weitz, K., André, E., & Rehm, M. (2021, November). "An Error Occurred!" Trust repair with virtual robot using levels of mistake explanation. *In Proceedings of the 9th International Conference on Human-Agent Interaction* (pp. 218–226). Virtual Event Japan.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517−527.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors, 63*(7), 1196−1229.

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016, August). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. *In 2016 25th IEEE international symposium on robot and human interactive communication*(RO-MAN) (pp. 493‒500). New York,

Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *In Proceedings of the world congress on engineering and computer science* (pp. 425–430).San Francisco, USA.

Haring, K. S., Satterfield, K. M., Tossell, C. C., De Visser, E. J., Lyons, J. R., Mancuso, V. F., ... & Funke, G. J. (2021). Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. *Frontiers in Psychology, 12*, 625713.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407−434.

Hopko, S. K., & Mehta, R. K. (2022). Trust in shared-space collaborative robots: Shedding light on the human brain. *Human Factors*, in press. https://doi.org/10.1177/00187208221109039.

Hou, Y. T. Y., & Jung, M. F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*, CSCW2, 477.

Jensen, T., Albayram, Y., Khan, M. M. H., Fahim, M. A. A., Buck, R., & Coman, E. (2019, June). The apple does fall far from the tree: User separation of a system from its developers in human-automation trust repair. *In Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 1071–1082). San Diego, CA, USA.

Jessup, S. A., Gibson, A., Capiola, A. A., Alarcon, G. M., & Borders, M. (2020, January). Investigating the effect

of trust manipulations on affect over time in human-human versus human-robot interactions. *Proceedings of the 53rd Hawaii International Conference on System Sciences*(pp. 1–10).

Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE*, 80–87.

Kaniarasu, P., & Steinfeld, A. M. (2014, August). Effects of blame on trust in human robot interaction. *In The 23rd IEEE international symposium on robot and human interactive communication* (pp. 850–855). Edinburgh, Scotland, UK.

Khavas, Z. R. (2021). A review on trust in human-robot interaction. *arXiv preprint arXiv:2105.10045.*

Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2020, November). Modeling trust in human-robot interaction: A survey. *In Social Robotics: 12th International Conference, ICSR* (pp. 529–541). https://doi.org/10.1007/978-3-030-62056-1_44.

Kim, D., & Kim, S. (2021). A model for user acceptance of robot journalism: Influence of positive disconfirmation and uncertainty avoidance. *Technological Forecasting and Social Change, 163*, 120448.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review, 34*, 401–422.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104−118.

Kim, T., & Hinds, P. (2006, September). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *In ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication* (pp. 80–85). Hatfield, UK.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics, 61*, 101595.

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems, 35*(2), 30.

Kraus, J., Scholz, D., Messner, E. M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology, 10*, 2917. doi: 10.3389/fpsyg.2019.02917.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors, 62*(5), 718–736.

Kundinger, T., Wintersberger, P., & Riener, A. (2019, May). (Over) Trust in automated driving: The sleeping pill of tomorrow? *In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). Glasgow, Scotland UK.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics, 62*(3), 345–360.

Kwon, J. H., Jung, S. H., Choi, H. J., & Kim, J. (2021). Antecedent factors that affect restaurant brand trust and brand loyalty: Focusing on US and Korean consumers. *Journal of Product & Brand Management, 30*(7), 990–1015.

Lee, J. D., & Kolodge, K. (2020). Exploring trust in self-driving vehicles through text analysis. *Human Factors, 62*(2), 260–277.

Lee, J.D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243−1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International*

Journal of Human-Computer Studies, 40(1), 153−184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50−80.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 1–16.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010, March). Gracefully mitigating breakdowns in robotic services. *In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 203–210).Osaka, Japan.

Lee, S. L., Lau, I. Y. M., Kiesler, S., & Chiu, C. Y. (2005, April). Human mental models of humanoid robots. *In Proceedings of the 2005 IEEE international conference on robotics and automation* (pp. 2767–2772). Barcelona, Spain.

Li, D., Rau, P. P., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics, 2*, 175−186.

Liu, X. S., Yi, X. S., & Wan, L. C. (2022). Friendly or competent? The effects of perception of robot appearance and service context on usage intention. *Annals of Tourism Research, 92*, 103324.

Löffler, D., Dörrenbächer, J., & Hassenzahl, M. (2020, March). The uncanny valley effect in zoomorphic robots: The U-shaped relation between animal likeness and likeability. *In Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 261–270). Cambridge, United Kingdom.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90−103.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423–431.

Lyons, J. B., Hamdan, I., & Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior, 138*, 107473.

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020, September). The role of individual differences as predictors of trust in autonomous security robots. *In 2020 IEEE International Conference on Human-Machine Systems (ICHMS)* (pp. 1–5). Rome, Italy.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., ... & Shively, R. (2017). Shaping trust through transparent design: theoretical and experimental guidelines.In: Savage-Knepshield, P., & Chen, J(Eds.), *Advances in Human Factors in Robots and Unmanned Systems*(pp.127–136). Springer International Publishing.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277−301.

Martinez, J. E., VanLeeuwen, D., Stringam, B. B., & Fraune, M. R. (2023, March). Hey? ! What did you think about that robot? Groups polarize users' acceptance and trust of food delivery robots. *In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 417–427). https://doi.org/10.1145/3568162.3576984.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48*(4), 656−665.

Meng, J., & Berger, B. K. (2019). The impact of organizational culture and leadership performance on PR professionals' job satisfaction: Testing the joint mediating effects of engagement and trust. *Public Relations Review, 45*(1), 64–75.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don' t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI, 4*, 21.

Montague, E., & Xu, J. (2012). Understanding active and passive users: The effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user. *Applied Ergonomics, 43*(4), 702–712.

Montague, E., Xu, J., & Chiou, E. (2014). Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. *IEEE Transactions on Human-Machine Systems, 44*(5), 614–624.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In Parasuraman, R., & Mouloua. M(Eds.), *Automation and Human Performance*(pp.201–220). CRC Press.

Müller, R., Schischke, D., Graf, B., & Antoni, C. H. (2023) How can we avoid information overload and techno-frustration as a virtual team? The effect of shared mental models of information and communication technology on information overload and techno-frustration. *Computers in Human Behavior, 138*, 107438.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies, 169*, 102941.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: When design meets trust calibration. *World Wide Web, 24*(5), 1857–1884.

Oh, S., Seong, Y., Yi, S., & Park, S. (2020). Neurological measurement of human trust in automation using electroencephalogram. *International Journal of Fuzzy Logic and Intelligent Systems, 20*(4), 261–271.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *Plos One, 15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132.

Okuoka, K., Enami, K., Kimoto, M., & Imai, M. (2022). Multi-device trust transfer: Can trust be transferred among multiple devices? *Frontiers in Psychology, 13*, 920844.

Onnasch, L., & Panayotidis, T. (2020, December). Social loafing with robots-An empirical investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(*1), 97–101.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building appropriate trust in human-robot teams. In 2013 AAAI spring symposium series.

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652.*

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381−410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230−253.

Perkins, R., Khavas, Z. R., & Robinette, P. (2021). Trust calibration and trust respect: A method for building team cohesion in human robot teams. *arXiv preprint arXiv:2110.06809.*

Petrocchi, S., Iannello, P., Lecciso, F., Levante, A., Antonietti, A., & Schulz, P. J. (2019). Interpersonal trust in doctor-patient relation: Evidence from dyadic analysis and association with quality of dyadic communication. *Social Science & Medicine, 235,* 112391.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57*(4), 545–556.

Pynadath, D. V., Wang, N., & Kamireddy, S. (2019, September). A Markovian Method for Predicting Trust

Behavior in Human-Agent Interaction. *In Proceedings of the 7th International Conference on Human-Agent Interaction* (pp. 171–178). Kyoto, Japan.

Quinn, D. B. (2018). *Exploring the efficacy of social trust repair in human-automation interactions* (Doctoral dissertation). Clemson University, Lawton.

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016, August). Errare humanum est: Erroneous robots in human-robot interaction. *In 2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)* (pp. 501–506). New York, NY, USA.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95–112.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. *In Social Robotics: 7th International Conference, ICSR,* Paris, France.

Robinette, P., Howard, A. M., & Wagner, A. R. (2017a). Conceptualizing overtrust in robots: Why do people trust a robot that previously failed?. In: Lawless, W., Mittu, R., Sofge, D., & Russell, S (Eds), *Autonomy and Artificial Intelligence: A Threat or Savior?*(pp. 129–155). Springer, Cham.

Robinette, P., Howard, A. M., & Wagner, A. R. (2017b). Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425–436.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. *In 2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 101–108). Christchurch, New Zealand.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2017, November). Human perceptions of the severity of domestic robot errors. *In Social Robotics: 9th International Conference(ICSR)*(pp. 647–656).Tsukuba, Japan.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics, 5*, 313−323.

Sanders, T. L., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors, 61*(4), 614−626.

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., & Hancock, P. A. (2017, September). Trust and prior experience in human-robot interaction. *In Proceedings of the human factors and ergonomics society annual meeting* (pp. 1809–1813). Sage CA: Los Angeles, CA.

Sarkar, S., Araiza-Illan, D., & Eder, K. (2017). Effects of faults, experience, and personality on trust in a robot co-worker. *arXiv preprint arXiv:1703.02335.*

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019, March). "I don't believe you": Investigating the effects of robot trust violation and repair. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 57–65). Daegu, Korea (South).

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics, 38*(7-8), 608−625.

Shank, D. B., Bowen, M., Burns, A., & Dew, M. (2021). Humans are perceived as better, but weaker, than artificial intelligence: A comparison of affective impressions of humans, AIs, and computer systems in roles on teams. *Computers in Human Behavior Reports, 3*, 100092.

Shi, Y., Azzolin, N., Picardi, A., Zhu, T., Bordegoni, M., & Caruso, G. (2020). A Virtual reality-based platform to validate HMI design for increasing user' s trust in autonomous vehicle. *Computer-Aided Design and Applications, 18*(3), 502–518.

Shin, D., Zaid, B., & Ibahrine, M. (2020, November). Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. *In 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)* (pp. 1–5). Sharjah, United Arab Emirates.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010, March). No fair! An interaction with a cheating robot. *In 2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 219–226). Osaka, Japan.

Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors, 20*(18), 5087.

Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation, 55*, 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

Tam, K. Y., & Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research, 16*(3), 271–291.

Toader, D. C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2019). The effect of social presence and chatbot errors on trust. *Sustainability, 12*(1), 256.

Ullman, D., & Malle, B. F. (2017, March). Human-robot trust: Just a button press away. *In Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 309–310). Vienna, Austria.

van Maris, A., Lehmann, H., Natale, L., & Grzyb, B.(2017, March). The influence of a robot' s embodiment on trust: A longitudinal study. *In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on human-robot interaction* (pp. 313–314). Vienna, Austria.

van Pinxteren, M. M., Wetzels, R. W., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: Implications for services marketing. *Journal of Services Marketing, 33*(4), 507–518.

Volante, W. G., Sosna, J., Kessler, T., Sanders, T., & Hancock, P. A. (2019). Social conformity effects on trust in simulation-based human-robot interaction. *Human Factors, 61*(5), 805–815.

Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM, 61*(9), 22–24.

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation research part F: Traffic psychology and behaviour, 64*, 401–412.

Wang, N., Pynadath, D.V., Rovira, E., Barnes, M.J., Hill, S.G. (2018). Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In: Ham, J., Karapanos, E., Morita, P., & Burns, C(Eds), *Persuasive Technology* (pp. 56–69). Springer, Cham.

Washburn, A., Adeleye, A., An, T., & Riek, L. D. (2020). Robot errors in proximate HRI: How functionality framing affects perceived reliability and trust. *ACM Transactions on Human-Robot Interaction (THRI), 9*(3), 1–21.

Wickens, C. D. (1995). Designing for situation awareness and trust in automation. *IFAC Proceedings Volumes, 28*(23), 365–370.

Wullenkord, R., Fraune, M. R., Eyssel, F., & Šabanović, S. (2016, August). Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. *In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN) (pp. 980–985). New York, USA.

Yen, C., & Chiang, M. C. (2021). Trust me, if you can: A study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behaviour & Information Technology, 40*(11), 1177–1194.

Xu, J., De'Aira, G. B., & Howard, A. (2018, August). Would you trust a robot therapist? Validating the equivalency of trust in human-robot healthcare scenarios. *In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 442–447). Nanjing, China.

Xu, J., & Howard, A. (2018, August). The impact of first impressions on human-robot trust during problem-solving scenarios. *In 2018 27th IEEE international symposium on robot and human interactive*

communication (RO-MAN) (pp. 435–441). Nanjing, China.
Xu, J., & Montague, E. (2013, September). Group polarization of trust in technology. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 344–348). Sage CA: Los Angeles, CA.

Trust dampening and trust promoting：A dual-pathway of trust calibration in human-robot interaction

HUANG Xinyu, LI Ye

(School of Psychology, Central China Normal University & Key Laboratory of Adolescent Cyberpsychology and Behavior, Ministry of Education, Wuhan 430079,China)

Abstract: Trust is the basis of successful human-robot interaction. However, humans do not always hold the appropriate level of trust in human-robot interaction, sometimes they may also fall into pitfalls: the trust bias, which contains both over-trust and under-trust. Trust bias can harm the human-robot interaction and so trust calibration is necessary. Trust calibration is often achieved through two ways: trust dampening and trust promoting. Trust dampening focuses on how to reduce the high level of trust in robots, while trust promoting focuses on how to improve the low level of trust in robots. For future directions, we suggest further optimize the measurement of methods. Besides, we also need to clarify the cognitive process and explore more boundary conditions. Finally, in order to boost human-robot collaboration, researchers are encouraged to explore personalized and specialized trust calibration strategies based on individual differences and further clarify the various reasons why trust bias occurs.

Keywords: trust calibration, trust bias, trust dampening, trust promoting, human-robot interaction